Data mining in Prenzlauer Berg

Airbnb Project

OUTLINE

1. Data Preparation & Exploration
2. Prediction
3. Classification
4. Clustering

CONTENT

1 DATA PREPARATION & EXPLORATION

DATA PREPARATION & EXPLORATION

Data Cleaning

Summary Statistics
& Data Visualization

Mapping

Wordcloud

DATA CLEANING

- Read data into local environment
- Filter the data that pertain Prenzlauer Berg neighborhood only
- Delete non-meaning variables
- Deal with missing values
  - Count NAs by columns
  - Delete NA rows
  - Calculate columns' NA percentage
- Change numeric variables to categorical

SUMMARY STATISTICS

- Q1: What is the mean price for apartments in 2018?
- Q2: What is the price standard deviation for apartments in 2018?
- Q3: What is the price range for apartments in 2018?
- Q4: What is the quantile for apartments in 2018? Compare it with quantile on all property types in all years.

Apartment ($)	2018	All Years
Mean	184.5	192.34
SD	75.25	71.63
Range	10-294	1-294
Quantile	25% 75% 160.5 241	25% 75% 165 247

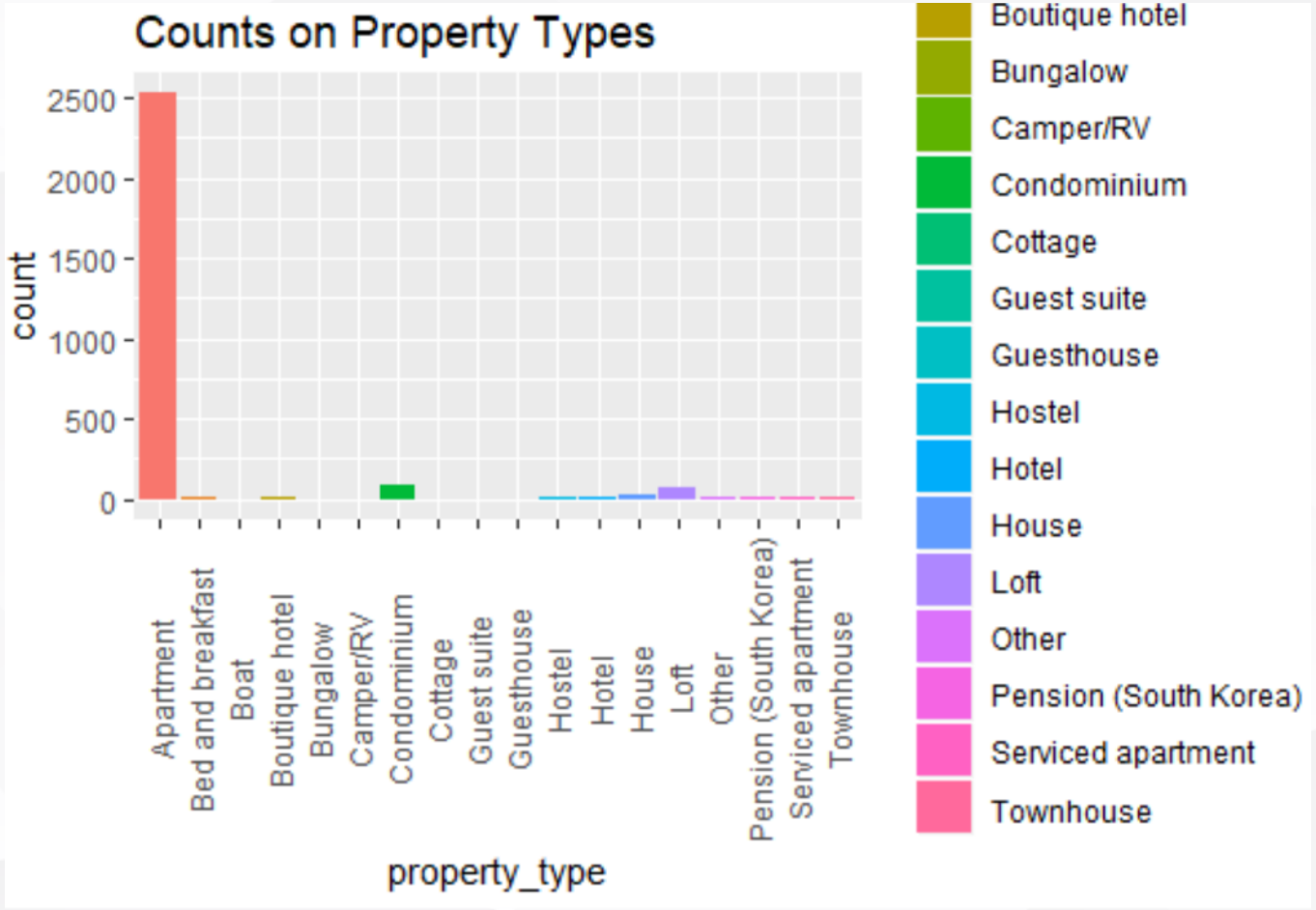
SUMMARY STATISTICS

- Q5: What is the percentage of each room type?

	room_type	n	per	label
	<fct>	<int>	<dbl>	<chr>
1	Shared room	21	0.00759	0.8%
2	Private room	1164	0.421	42.1%
3	Entire home/apt	1583	0.572	57.2%

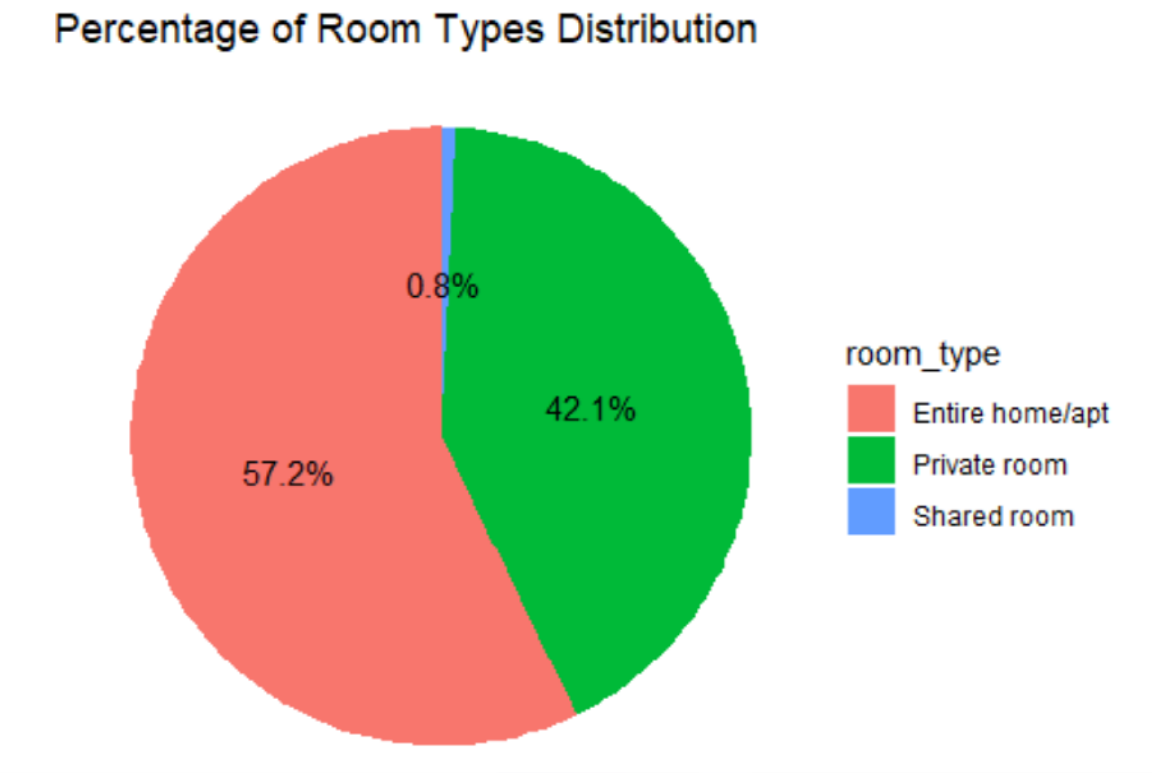
Highest

DATA VISUALIZATION

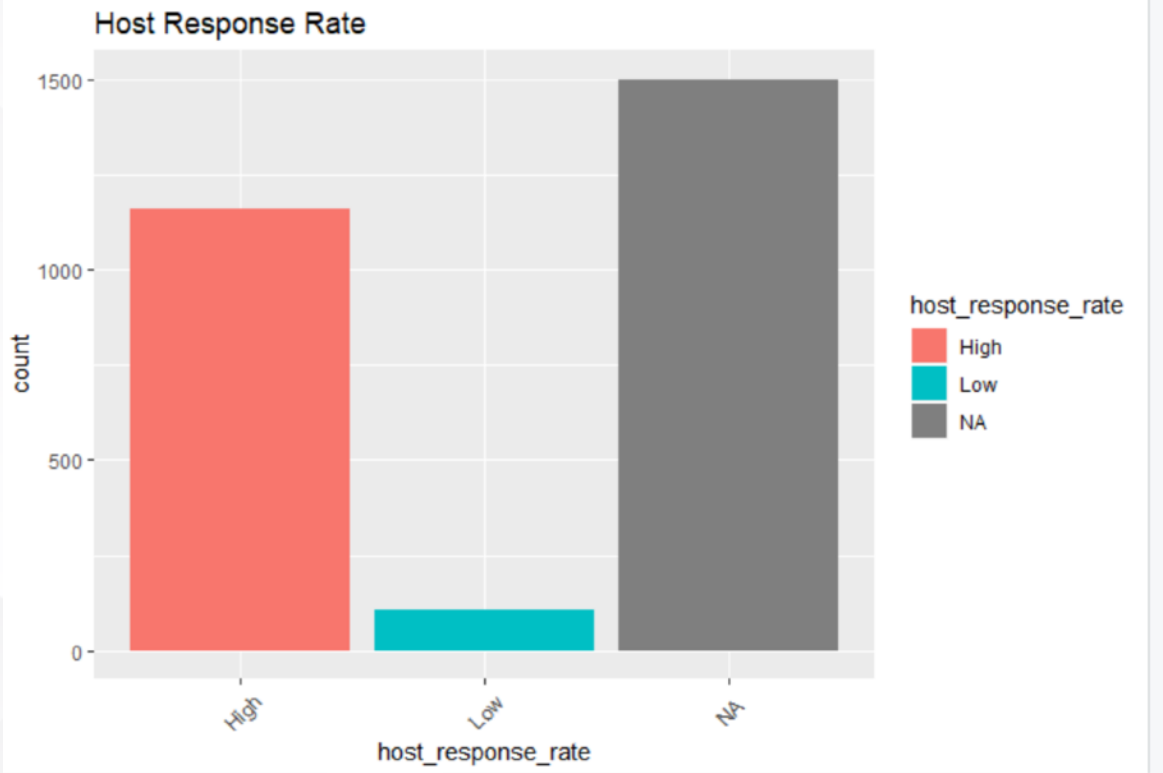


The bar chart can clearly show what property types available in this area and numbers for each of them.
It shows the majority property type is apartment; condominium rank as the second; loft rank as the third.

DATA VISUALIZATION

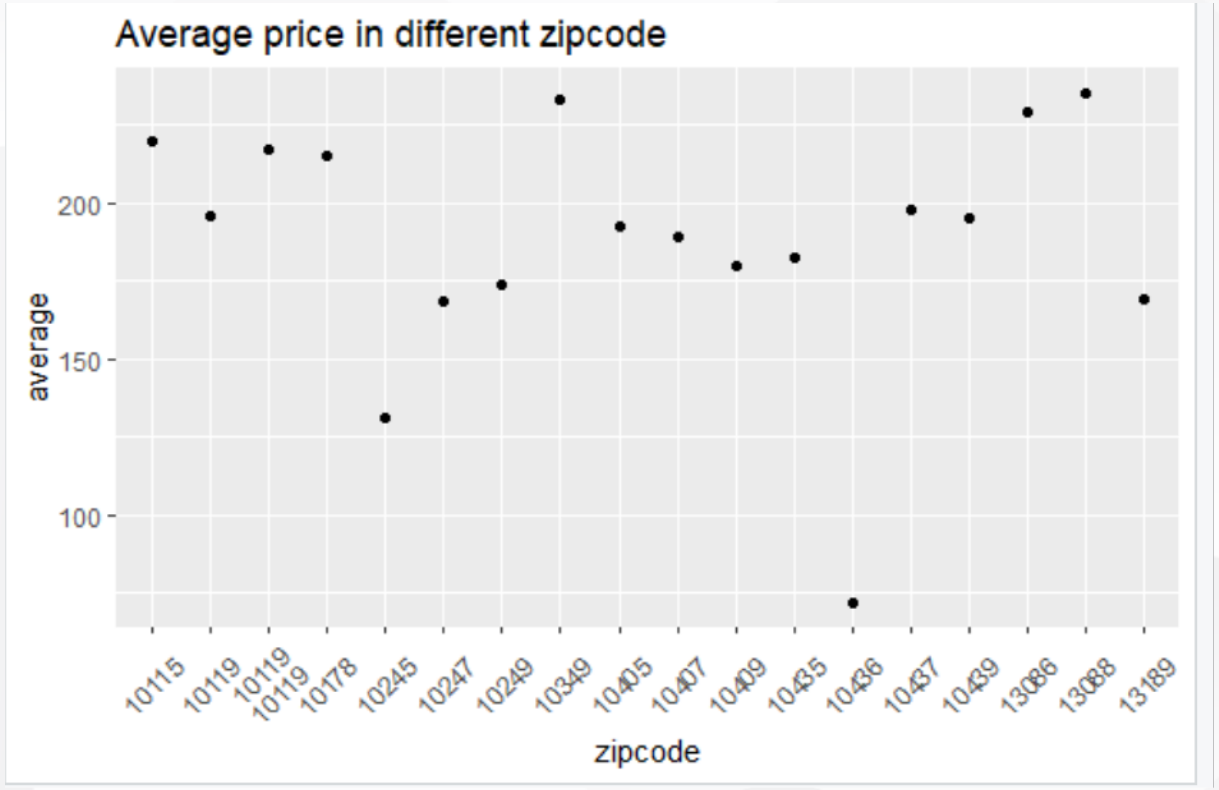


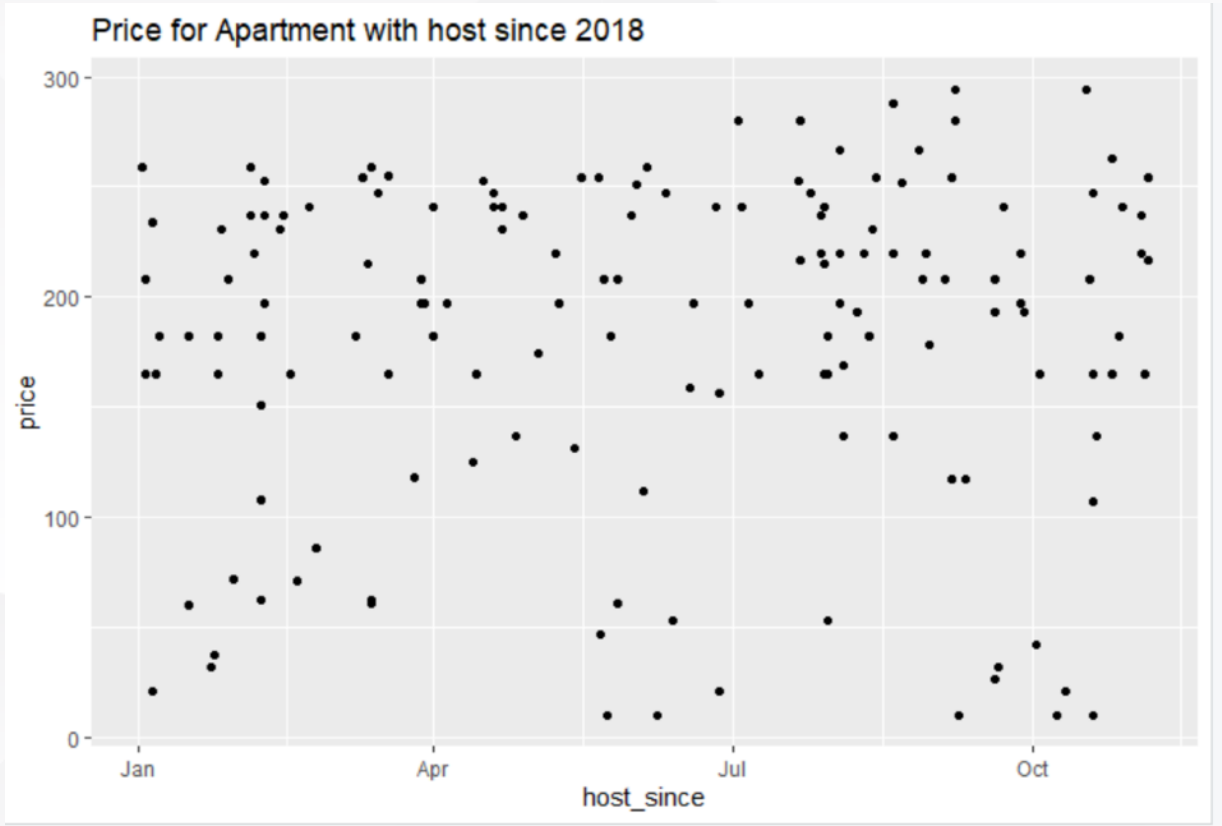
The pie chart shows the percentage distribution for each type



Regarding the response rate over 70 percent as high. The bar chart on host response rate helps to see the overall host response performance.

DATA VISUALIZATION

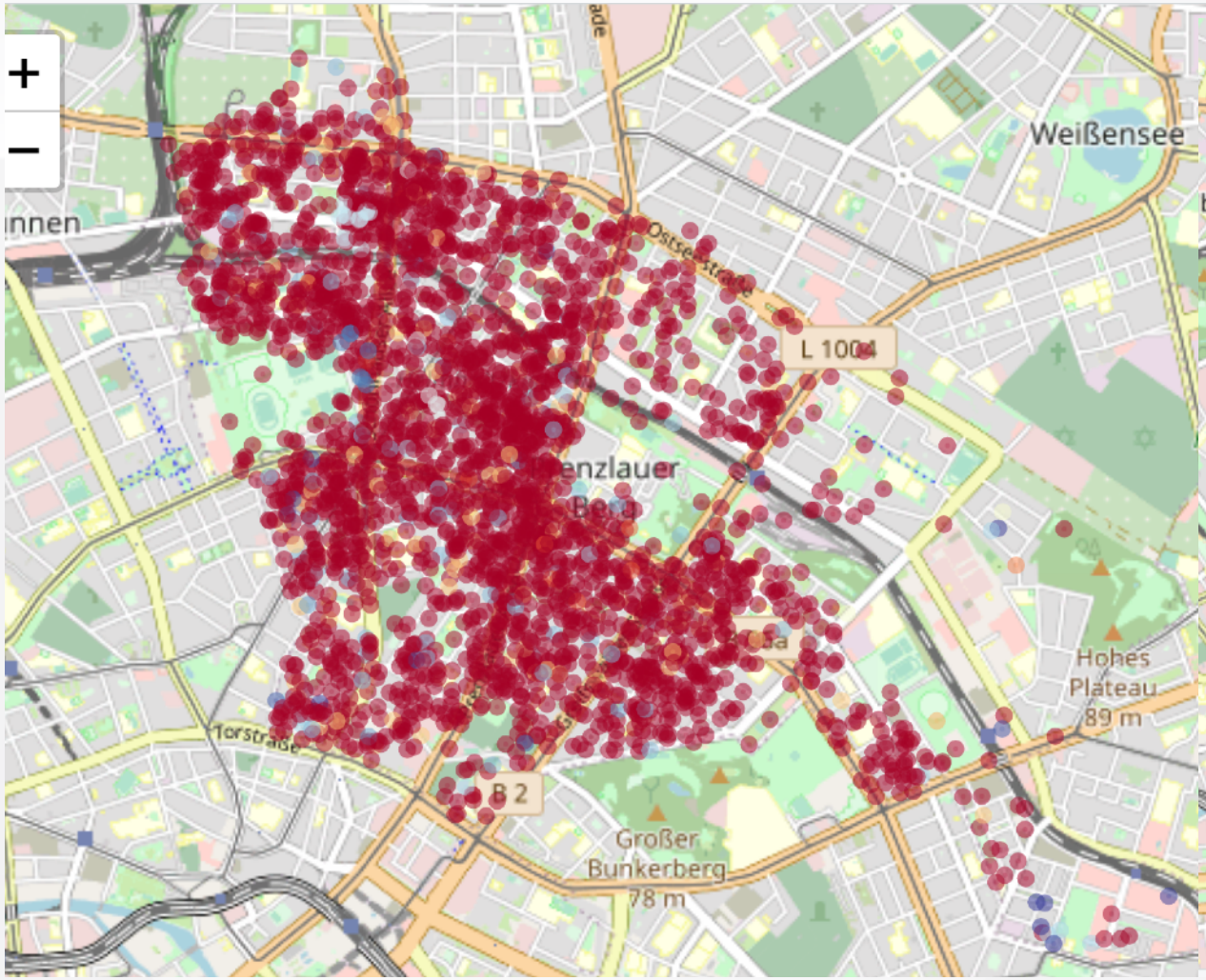


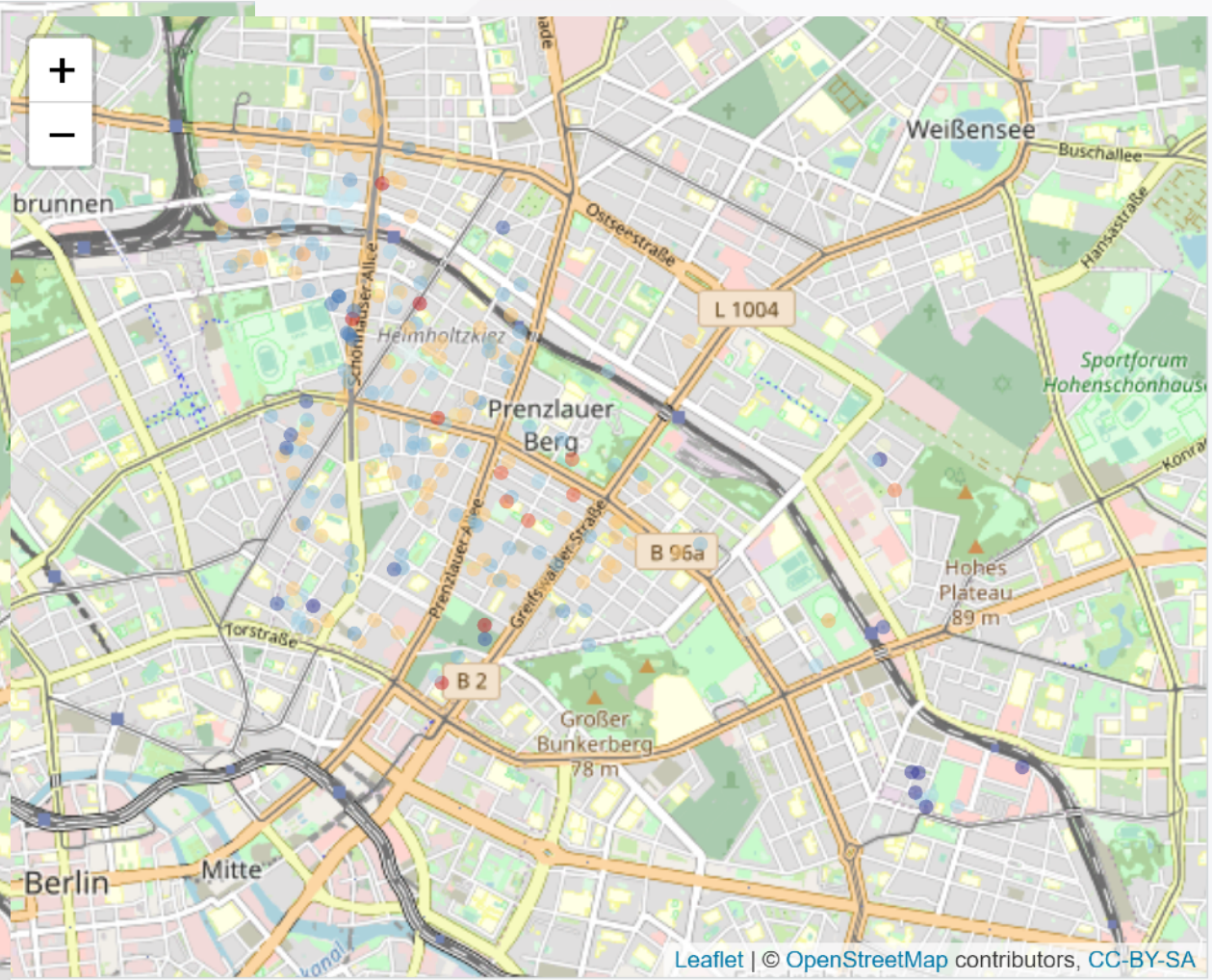


Prenzlauer Berg has different distinct areas that have different zip codes.
The scatter plot shows the average price in different zip area

Price distribution of apartment with host since 2018.

MAPPING


Weißensee
L 1004
Hohes Plateau 89 m
B 2
Großer Bunkerberg 78 m


Weißensee
Buschallee
Hansastraße
Ostseestraße
L 1004
Prenzlauer Berg
Sportforum Hohenschönhausen
B 96a
Hohes Plateau 89 m
Torstraße
B 2
Großer Bunkerberg 78 m
Berlin
Mitte
Leaflet | © OpenStreetMap contributors, CC-BY-SA

WORDCLOUD

The Word Cloud illustrates that when hosts describe the neighborhood, they pay much attention to the restaurants and shops around and whether the area is convenient for access to various dining or entertainment places.

CONTENT

2 PREDICTION

PREDICTION

Feature Selection

Before Modeling

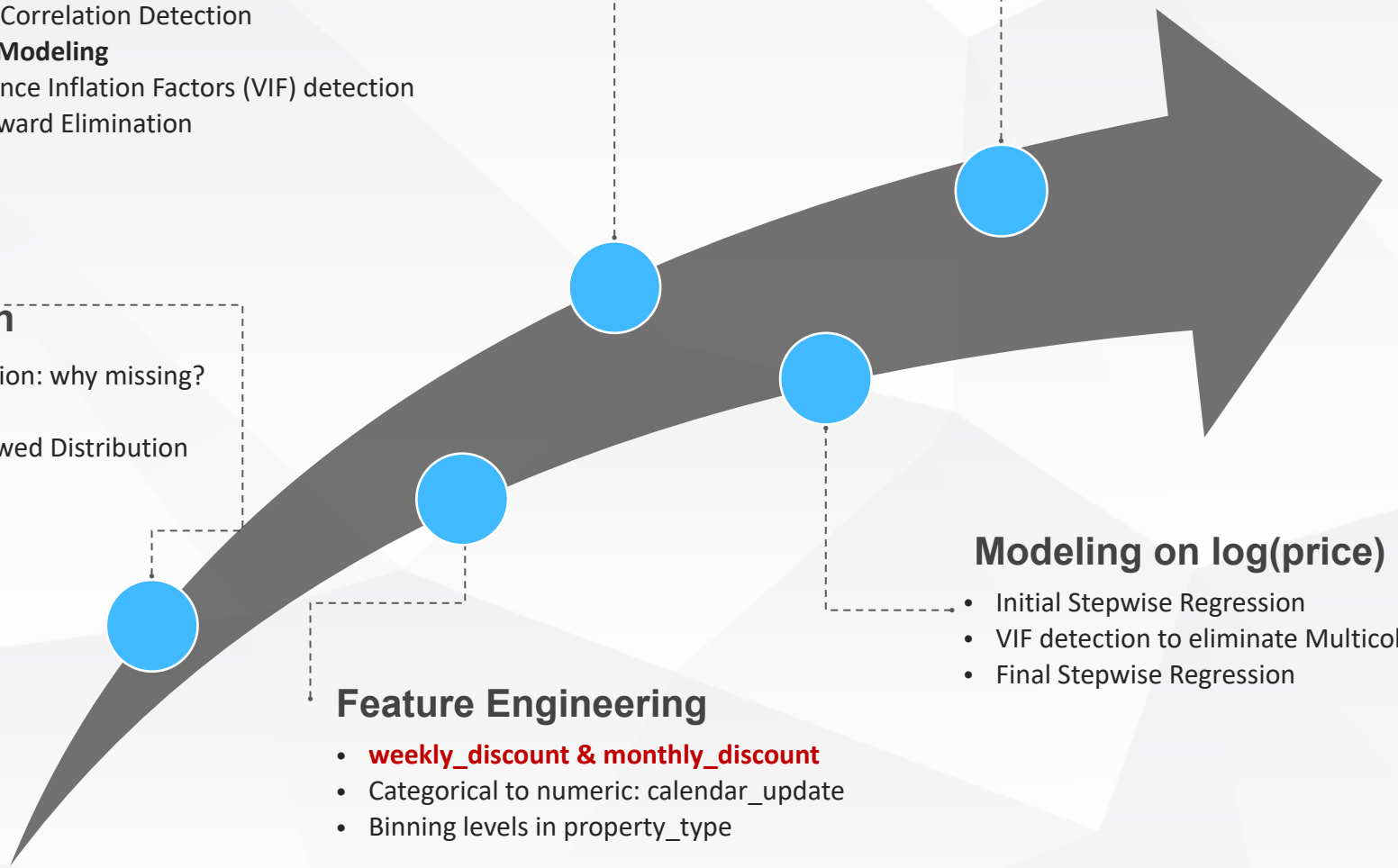
- Drop useless variables
- Delete the feature with high primary value
- High Correlation Detection

During Modeling

- Variance Inflation Factors (VIF) detection
- Backward Elimination

Performance Measurements

- **Adjusted R squared**: after log transformation on price, R squared improve from 0.04 to 0.52.
- **RMSE**: Training RMSE and Test RMSE is closed, so the model is good-fit and has generalization ability.
- **Residual Analysis** to check the assumptions of residual validate or not



Should Hosts Need to **Improve Cleaning**?

Data Preparation

- Missing Value Imputation: why missing?
- Data Transformation:
  Price follows right-skewed Distribution
  price => log(price)

Feature Engineering

- **weekly_discount & monthly_discount**
- Categorical to numeric: calendar_update
- Binning levels in property_type

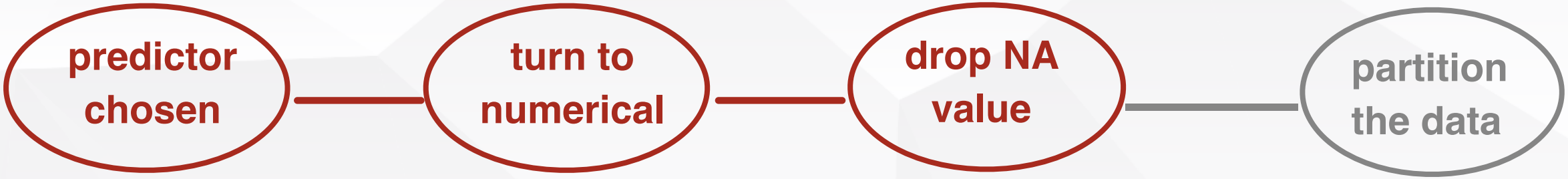
Modeling on log(price)

- Initial Stepwise Regression
- VIF detection to eliminate Multicollinearity problem
- Final Stepwise Regression

CONTENT

3 CLASSIFICATION

k-nearest neighbors



decided by the accuracy of k value

third try to gain largest k value

decided by the accuracy of k value

decided by the accuracy of k value

test the accuracy of choosing k value

- Reason for factor chosen
  - Affect the cancellation policy
  - involve the benefit of cleaning company

Find Best K Value & Most Accurate Model

FIRST TRY

```
> accuracy
    k  accuracy
1   1 0.4390244
2   2 0.4354110
3   3 0.4543812
4   4 0.4760614
5   5 0.4507678
6   6 0.4570912
7   7 0.4598013
8   8 0.45618
9   9 0.45618
10 10 0.4552846
11 11 0.4570912
12 12 0.4652213
13
14
```

SECOND TRY

```
> accuracy2 #smaller
    k accuracy2
1   1 0.3983740
2   2
3   3
4   4
5   5 0.3983740
6   6 0.3974706
7   7 0.4065041
8   8 0.4028907
9   9 0.4046974
10 10 0.3974706
11 11 0.4010840
12 12 0.4010840
```

THIRD TRY

```
> accuracy3
    k accuracy3
1   1 0.4333636
2   2 0.4451496
3   3 0.4660018
4   4 0.4723481
5   5 0.4759746
6   6 0.4714415
7   7 0.4723481
11 11 0.4877607
12 12 0.4841342
```

BEST ONE

Lorem Ipsum

Lorem Ipsum

Lorem Ipsum

Lorem Ipsum

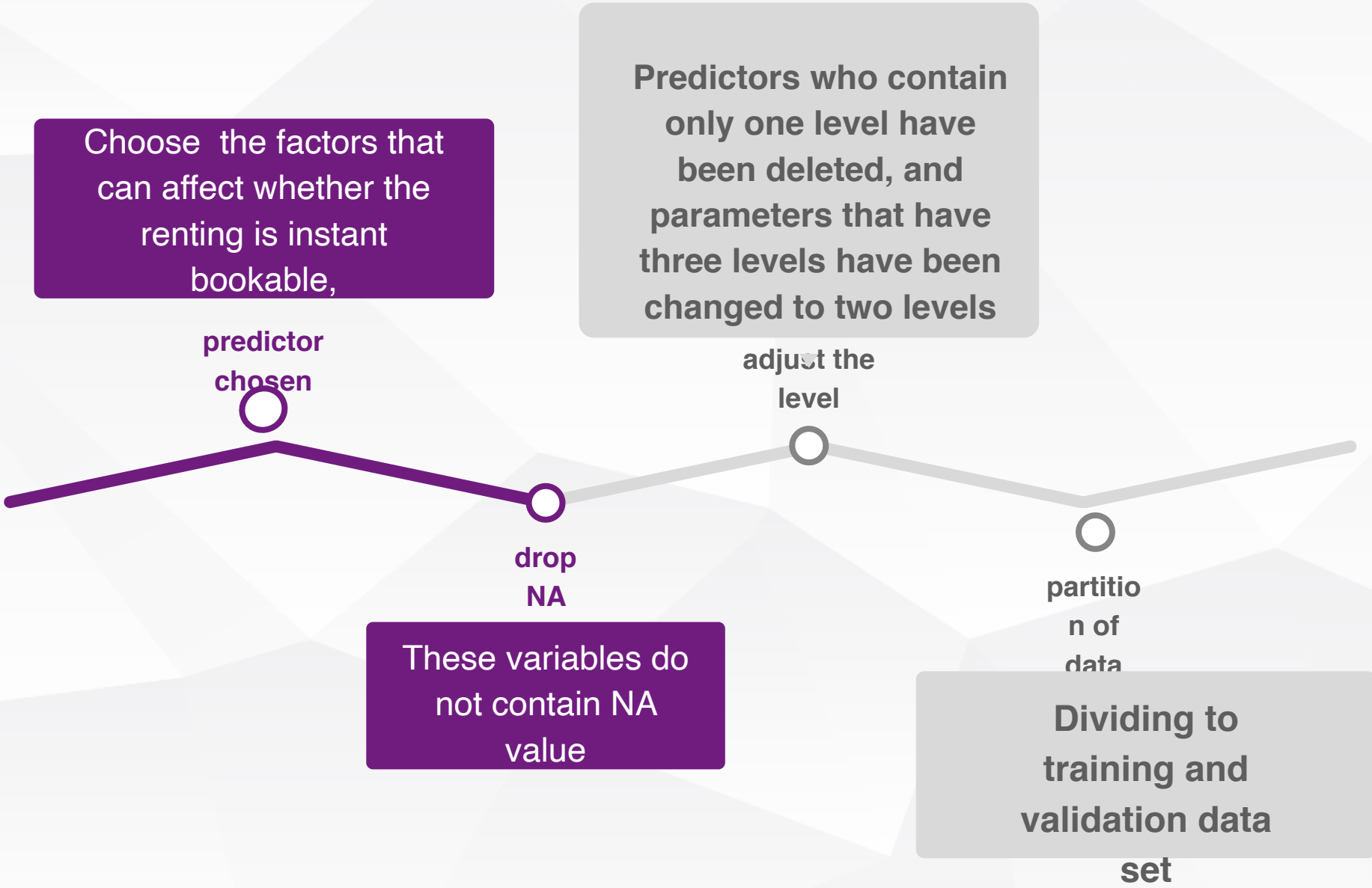
Lorem Ipsum

Lorem Ipsum

```
#third try#
PB_C3<- select(PB_2, id, accommodates, bathrooms, bedrooms, beds, price, cleaning_fee, guests_included, extra_people,
               minimum_nights, maximum_nights, number_of_reviews, cancellation_policy)
```

```
 [1] "878"  "1609" "539"  "200"  "1570" "279"  "81"   "1139" "1153" "128"  "582"  "468"
[13] "334"
> nn
[1] strict_14_with_grace_period
```

Naive Bayes


Choose the factors that can affect whether the renting is instant bookable,
predictor chosen
drop NA
These variables do not contain NA value
Predictors who contain only one level have been deleted, and parameters that have three levels have been changed to two levels
adjust the level
partitio n of data
Dividing to training and validation data set

Outcome of Naive Bayes

```
> answer
[1] f
Levels: f t
> predict(nbmodel,royal, type="raw")
              f         t
[1,] 0.6861451 0.3138549
```

```
Confusion Matrix and Statistics

          Reference
Prediction    f    t
         f 1198  459
         t    1    2

               Accuracy : 0.7229
                 95% CI : (0.7007, 0.7443)
    No Information Rate : 0.7223
    P-Value [Acc > NIR] : 0.4907
```

```
Confusion Matrix and Statistics

          Reference
Prediction   f   t
         f 776 330
         t   1   0

               Accuracy : 0.701
                 95% CI : (0.6731, 0.7279)
    No Information Rate : 0.7019
    P-Value [Acc > NIR] : 0.541

                  Kappa : -0.0018

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9987
            Specificity : 0.0000
         Pos Pred Value : 0.7016
         Neg Pred Value : 0.0000
             Prevalence : 0.7019
         Detection Rate : 0.7010
   Detection Prevalence : 0.9991
      Balanced Accuracy : 0.4994

       'Positive' Class : f
```

Classification Tree

01

Data leaning and Categorical or numeric data

02

Separates them into group

03

Divide data set into training and validation parts

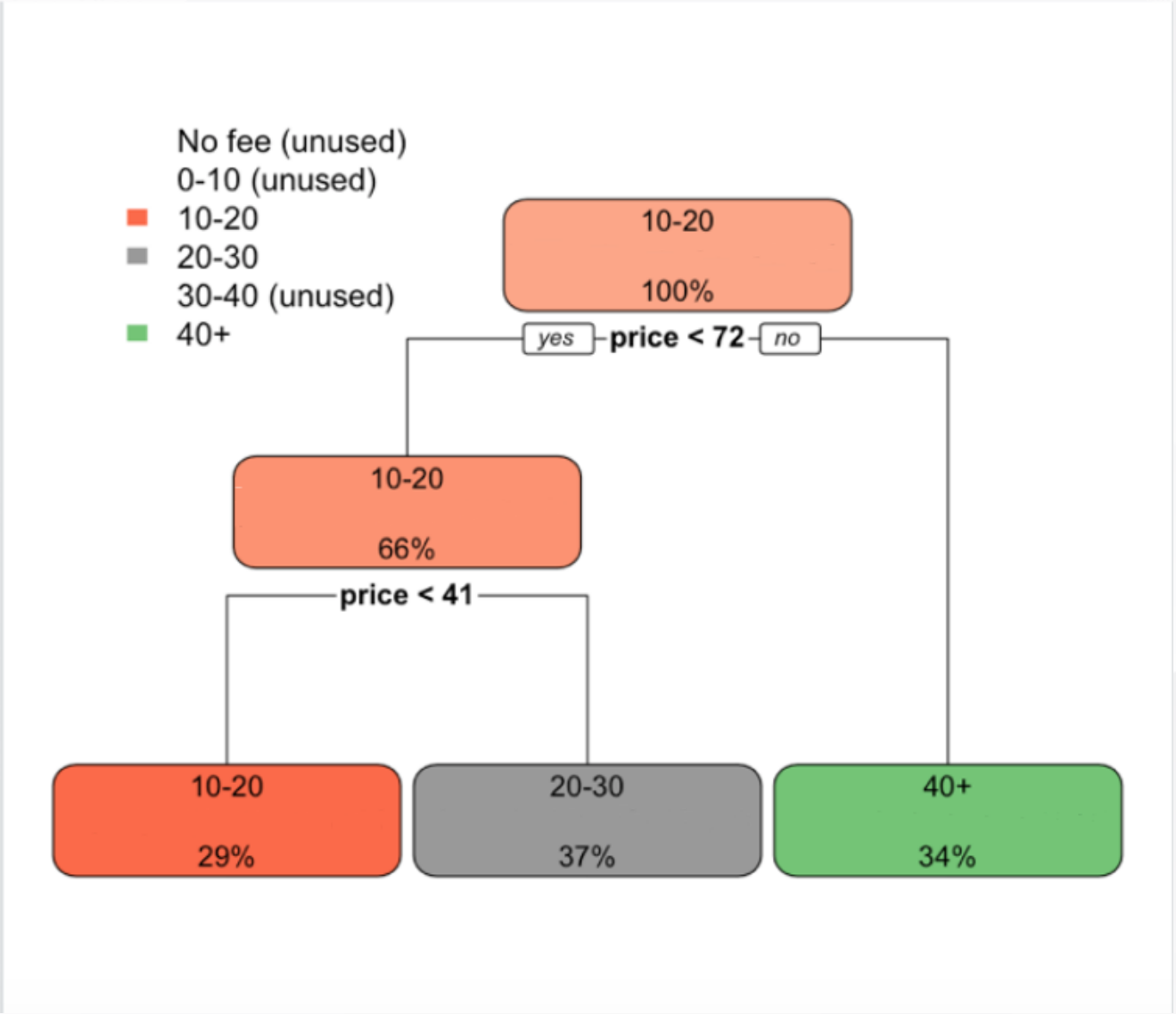
04

Rpart for running trees, Rpart.plot to plot them

05

Cross validation

Classification Tree



CONTENT

4 CLUSTERING

Data preprocess

Remove variables	**Square feet**
Replace NA	**By 0 : deposit, cleaning fee** **By median: review score, response rate**
Create new features	**Business interval = today – first review** **No Business interval = today – last review** **Operation interval = today – host since** **Weekly discount ratio = weekly price/(daily price*7)** **Monthly discount ratio = monthly price/(daily price*30)**

Cluster 1: discount rental

Features:

- Highest discount ratio provided to weekly or monthly rent.
- Highest score in customer review.
- Indicating long expectation of stay and high service level.

Ideal customers:

- Customers who are considering a weekly or monthly holiday.

weekly_discount_ratio	monthly_discount_ratio
0.18900439	0.3157696
-0.88110422	-1.5290928
0.00839183	0.0965972
0.06383194	0.1053612

review_scores_communication	review_scores_location	review_scores_value
0.22744756	0.18006151	0.3109978
0.02161257	-0.01939396	-0.1191145
-1.56561238	-1.11738084	-1.6194738
0.10802423	0.03288968	-0.0847248

review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin
0.276805691	0.26144594	0.2103013	0.23145905
0.003794764	0.01943824	0.1056233	0.03424158
-1.868786674	-1.73027722	-1.6672062	-1.59316842
0.127974343	0.07798181	0.1733943	0.09688596

Cluster 2: private room rental

Features:

Price is close to average (normal)

Highest in review per month and number of reviews.

Limited discount for long-term stay

Indicating frequently check-in and out of customer.

Ideal customers:

Traveler for short-term holiday.

price	security_deposit	cleaning_fee
-0.277889778	-0.13064716	-0.2474763
0.005597724	0.28061586	0.2915196
-0.259480483	-0.06673478	-0.191885
1.460373099	0.35013151	0.9670427

reviews_per_month	number_of_reviews
-0.07503059	-0.24366277
0.3446856	1.13559925
-0.25281874	-0.22840036
0.16684876	0.08093544

weekly_discount_ratio	monthly_discount_ratio
0.18900439	0.3157696
-0.88110422	-1.5290928
0.00839183	0.0965972
0.06383194	0.1053612

Cluster 3: shared room rental

Features:

- Second lowest in price
- Lowest in customer review
- Indicating a low price but low service level

Ideal customers:

- Backpackers
- Traveler with low budget in accommodation

price
-0.277889778
0.005597724
-0.259480483
1.460373099

review_scores_communication	review_scores_location	review_scores_value
0.22744756	0.18006151	0.3109978
0.02161257	-0.01939396	-0.1191145
-1.56561238	-1.11738084	-1.6194738
0.10802423	0.03288968	-0.0847248

review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin
0.276805691	0.26144594	0.2103013	0.23145905
0.003794764	0.01943824	0.1056233	0.03424158
-1.868786674	-1.73027722	-1.6672062	-1.59316842
0.127974343	0.07798181	0.1733943	0.09688596

Cluster 4: entire house rental

Features:

- Highest price.
- Highest housing capacity – beds, bathrooms.
- High deposit and cleaning fee required
- Highest number of guest and extra people
- Indicating big space rental

accommodates	bathrooms	bedrooms	beds
-0.32812789	-0.1231097	-0.2772093	-0.31061762
-0.08405867	-0.1042527	-0.1792372	-0.1446121
-0.08930159	-0.1323639	-0.1722771	-0.08981272
1.65810346	0.7732995	1.5881741	1.64260942

Ideal customers:

- Big group size: family or community

price	security_deposit	cleaning_fee	guests_included	extra_people
-0.277889778	-0.13064716	-0.2474763	-0.24012952	-0.1596507
0.005597724	0.28061586	0.2915196	-0.02440358	0.2480497
-0.259480483	-0.06673478	-0.191885	-0.09434024	-0.1325878
1.460373099	0.35013151	0.9670427	1.19573591	0.5667669

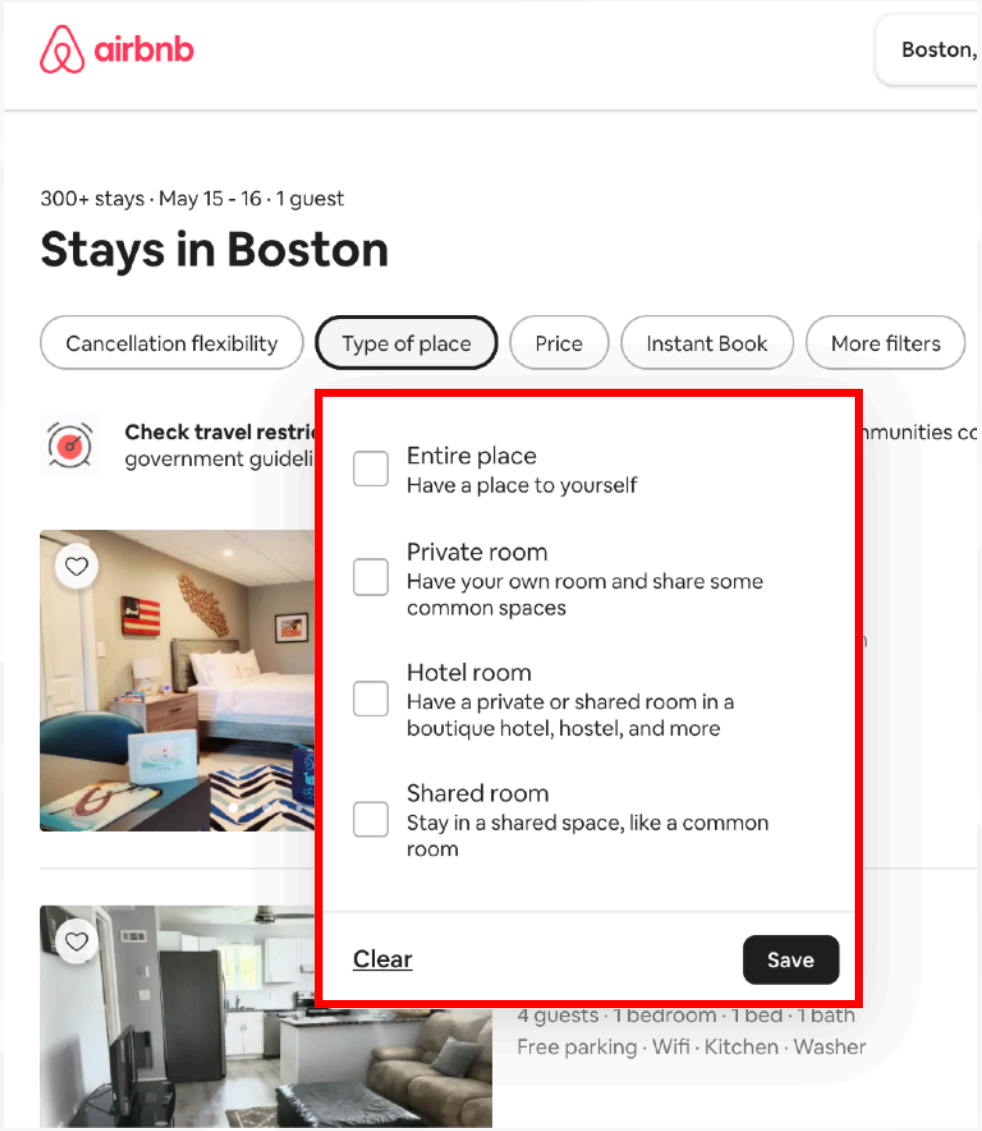
Clustering application

Website filter

Allow users to easily find out the place they prefer based on their preferences.

Host prioritization

Enable Airbnb to notice what kind of host are matter to them, so it can develop different strategies to maintain client relationship.



THANK YOU!

Any Questions?